Complex NMF under spectrogram consistency constraints *

◎ Jonathan Le Roux, Hirokazu Kameoka (NTT CS Labs),
Emmanuel Vincent (INRIA), Nobutaka Ono (The University of Tokyo),
Kunio Kashino (NTT CS Labs) and Shigeki Sagayama (The University of Tokyo)

1 Introduction

Many audio signal processing algorithms rely on the estimation of magnitude or complex short-time Fourier transform (STFT) spectrograms, but usually do not take into account the necessity for the estimated spectrograms to be consistent, i.e., to correspond to the STFT of a real-valued time-domain signal. Consistency constraints were introduced in [1] and applied there to phase reconstruction from magnitude spectrograms. In this paper, we show how to use them to introduce penalty functions on the consistency of STFT spectrograms into the recently introduced complex non-negative matrix factorization (NMF) framework [2], which estimates recurring patterns in the observed magnitude spectra, their activations and their phases. We derive analytical update equations through an auxiliary function approach, and present preliminary results on a supervised monaural source separation task.

2 Presentation of the model

The complex NMF model is a mixture model defined in the complex time-frequency domain. We assume that the modeled spectrogram $F_{\omega,t}$ in frequency bin ω and time frame t is the sum of K component spectrograms $W^k_{\omega,t}$ expressed as

$$W^k_{\omega,t} = H^k_\omega U^k_t e^{j\phi^k_{\omega,t}}, \tag{1}$$

where $H^k_\omega \geq 0$ corresponds to recurring magnitude spectral patterns, $U^k_t \geq 0$ to time-varying activation coefficients and $\phi^k_{\omega,t}$ to time-varying phase spectra. H^k_ω is normalized to avoid scaling ambiguities: $\forall k, \sum_\omega H^k_\omega = 1$. The problem is now, given an observed spectrogram $Y_{\omega,t}$, to estimate the optimal parameters $\theta = \{H, U, \phi\}$ of the model. It was formulated in [2] as the minimization of the L^2 norm between the observation and the model with a generalized Gaussian prior on U to promote sparsity.

We further introduce consistency penalty functions on each W^k. Let w and s be analysis and synthesis windows of length N verifying the perfect reconstruction conditions $1 = \sum_{q=0}^{Q-1} w(t-qR)s(t-qR)$, $\forall t$, for a frame shift R, where $Q = N/R$. One can show [1] that the set of consistent spectrograms is the kernel of the $\mathbb{R}$-linear operator from $\mathbb{C}^{MN}$ (M denoting the number of frames) to itself defined by

$$\mathcal{F}_{w,s}(W) = (\mathrm{STFT}_w \circ \mathrm{iSTFT}_s - \mathbf{I}_{MN})(W). \tag{2}$$

We use the L^2 norm of $\mathcal{F}_{w,s}(W^k)$ as a penalty to promote consistency on each separated spectrogram. The problem becomes that of minimizing

$$f(\theta) = ||Y - F||^2 + 2\lambda \sum_{k,t} |U^k_t|^p + \gamma \sum_k ||\mathcal{F}(W^k)||^2,$$

where p is a shape parameter which promotes sparsity for $0 < p < 2$, and λ and γ are prior weights.

3 Optimization

The optimization of the complex NMF model parameters was performed in [2] through an efficient iterative algorithm based on an auxiliary function approach. We derive here an auxiliary function for the new consistency term.

Let $(A^{\omega,t}_{\omega',t'})$ be the matrix representation of $\mathcal{F}$. We then have $\mathcal{F}(W)_{\omega,t} = \sum_{\omega',t'} A^{\omega,t}_{\omega',t'} W_{\omega',t'}$. For any auxiliary variables $\bar{Z}^{\omega,t,k}_{\omega',t'}$, $\bar{Y}^k$, $\bar{U}^k_t$ s.t. $\forall \omega,t,k$, $\sum_{\omega',t'} \bar{Z}^{\omega,t,k}_{\omega',t'} = 0$, $\sum_k \bar{Y}^k_{\omega,t} = Y_{\omega,t}$, $\bar{U}^k_t \in \mathbb{R}$, and for any $\beta^k_{\omega,t} > 0$, $\delta^{\omega,t,k}_{\omega',t'} > 0$ s.t. $\forall \omega,t,k$, $\sum_{\omega',t'} \delta^{\omega,t,k}_{\omega',t'} = 1$, $\sum_k \beta^k_{\omega,t} = 1$, we can show that $f(\theta) \leq f^+(\theta, \bar{\theta})$ with the auxiliary function f^+ defined as

$$\begin{aligned} f^+(\theta,\bar{\theta}) = &\sum_{k,\omega,t} \frac{|\bar{Y}^k_{\omega,t} - H^k_\omega U^k_t e^{j\phi^k_{\omega,t}}|^2}{\beta^k_{\omega,t}} \\ &+ \lambda \sum_{k,t} \big(p|\bar{U}^k_t|^{p-2}(U^k_t)^2 + (2-p)|\bar{U}^k_t|^p\big) \\ &+ \gamma \sum_{k,\omega,t,\omega',t'} \frac{1}{\delta^{\omega,t,k}_{\omega',t'}} \Big| \bar{Z}^{\omega,t,k}_{\omega',t'} - A^{\omega,t}_{\omega',t'} H^k_{\omega'} U^k_{t'} e^{j\phi^k_{\omega',t'}} \Big|^2, \end{aligned} \tag{3}$$

and $\bar{\theta} = \{\bar{Y}, \bar{U}, \bar{Z}\}$. f^+ is minimized w.r.t. $\bar{\theta}$ when

$$\begin{aligned} \bar{Y}^k_{\omega,t} &= H^k_\omega U^k_t e^{j\phi^k_{\omega,t}} + \beta^k_{\omega,t}(Y_{\omega,t} - F_{\omega,t}) \\ \bar{U}^k_t &= U^k_t \\ \bar{Z}^{\omega,t,k}_{\omega',t'} &= A^{\omega,t}_{\omega',t'} H^k_{\omega'} U^k_{t'} e^{j\phi^k_{\omega',t'}} - \delta^{\omega,t,k}_{\omega',t'} (\mathcal{F}(W^k))_{\omega,t}. \end{aligned}$$

The update equation for ϕ can be obtained as

$$\phi^k_{\omega,t} \leftarrow \mathrm{Arg}\Big(\frac{\bar{Y}^k_{\omega,t}}{\beta^k_{\omega,t}} + \gamma\big(a^k_{\omega,t} W^k_{\omega,t} - (\mathcal{F}^H \mathcal{F}(W^k))_{\omega,t}\big)\Big) \tag{4}$$

where $a^k_{\omega',t'} = \sum_{\omega,t} |A^{\omega,t}_{\omega',t'}|^2 / \delta^{\omega,t,k}_{\omega',t'}$ and $\mathcal{F}^H$ denotes the Hermitian adjoint of $\mathcal{F}$, which can be computed very efficiently by noticing that $\mathcal{F}^H_{w,s} = \mathcal{F}_{s,w}$.

If we assume that first $\bar{Y}$ then $\bar{U}$ then ϕ have been updated, and if we note $\bar{X}^k$ the term inside the argument in Eq. (4), then the updates for H and U become:

$$H^k_\omega \leftarrow \frac{\mathrm{Re}\big[\sum_t |\bar{X}^k_{\omega,t}| U^k_t\big]}{\sum_t (\frac{1}{\beta^k_{\omega,t}} + \gamma a^k_{\omega,t}) |U^k_t|^2}, \tag{5}$$

$$U^k_t \leftarrow \frac{\mathrm{Re}\big[\sum_\omega |\bar{X}^k_{\omega,t}| H^k_\omega\big]}{\sum_\omega (\frac{1}{\beta^k_{\omega,t}} + \gamma a^k_{\omega,t}) |H^k_\omega|^2 + \lambda p |\bar{U}^k_t|^{p-2}}. \tag{6}$$

---

* 無矛盾性拘束付き複素 NMF、ルルー・ジョナトン、亀岡弘和 (NTT)、ヴィンセント・エマヌエル (INRIA/IRISA)、小野順貴 (東大情報理工)、柏野邦夫 (NTT)、嵯峨山茂樹 (東大情報理工)

As in [2], β^k is set to $|W^k|/\sum_n |W^n|$. In order for the update equations to be tractable, we need to avoid the direct computation of δ when computing $a^k_{\omega,t}$. If we consider $\delta^{\omega,t,k}_{\omega',t'} = |A^{\omega,t}_{\omega',t'}|^q / \sum_{\omega',t'} |A^{\omega,t}_{\omega',t'}|^q$ where $q > 0$ is a tunable exponent, and notice [1] that $|A^{\omega,t}_{\omega',t'}| = |\alpha(\omega-\omega', t-t')|$ where the coefficients α depend on the windows w and s, then

$$a^k_{\omega,t} = \sum_{\omega',t'} |\alpha(\omega',t')|^{2-q} \sum_{\omega',t'} |\alpha(\omega',t')|^q = a. \quad (7)$$

Intuitively, a acts as a learning weight in Eq. (4): the larger a, the slower ϕ will move from its current value. As convergence is guaranteed anyway, we should thus look for a as small as possible, which is the case for $q = 1$, where we have $a = (\sum |\alpha|)^2$.

4 Phase reconstruction

Phase reconstruction with a given magnitude M can be considered as a particular case of the present framework. With $W = Me^{j\phi}$, minimizing $||\mathcal{F}(W)||^2$ gives the following update equation for ϕ:

$$\phi \leftarrow \mathrm{Arg}\big(aW - \mathcal{F}^H\mathcal{F}(W)\big). \quad (8)$$

It is interesting to note the link with the classical update by Griffin and Lim [3], which can be written

$$\phi \leftarrow \mathrm{Arg}\big(W + \mathcal{F}(W)\big), \quad (9)$$

If $w = s$ (e.g., for the sine window), then $\mathcal{F}^H = \mathcal{F}$ and one can see that $\mathcal{F}^H\mathcal{F}(W) = -\mathcal{F}(W)$. The only difference is then the a factor. We have not been able so far to find a setting for δ leading to a equal to or close to 1, and noticed through experiments that, due to that factor, Griffin and Lim's update was faster than the auxiliary approach one in terms of the decrease of inconsistency per iteration. We plan to investigate this issue in the future.

5 Experimental evaluation

We illustrate our method on a supervised speech-music source separation task in monaural conditions considered by Smaragdis [4]. Spectral bases are pre-trained on data from various sound classes with different spectral properties, fixed and used on mixtures of sounds from different classes to separate them. We use here chime sounds and speech uttered by a male speaker from the TIMIT database [4].

We trained 20 bases with classical NMF on 10 s of chime and speech data respectively. The sampling frequency was 16 kHz, and the spectrograms were built with a 32 ms length sine window and a 16 ms frame shift. We then concatenated the two bases and used 5 s of other parts of the chime and speech data to create a 0 dB mixture on which we tested our method, the original complex NMF without consistency constraints and the classical NMF. The goal of these experiments was to determine whether the consistency constraints helped improve the performance in a very simple setting. The sparsity parameters were set as in [2], the consistency parameter γ to 1 and a to $(\sum |\alpha|)^2 \approx 4.565$. U was initialized randomly, $e^{j\phi^k_{\omega,t}}$ initially set to $Y_{\omega,t}/|Y_{\omega,t}|$ and H fixed to the concatenation of the trained bases.

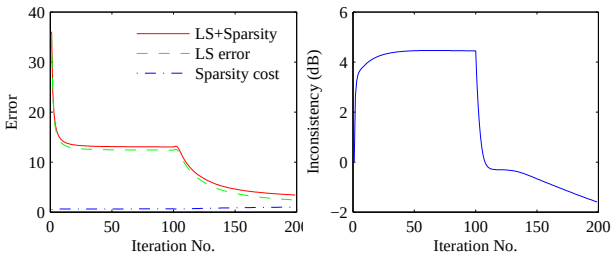

Fig. 1 Evolution of the components of the objective function: least-square and sparsity error (left), inconsistency (right).

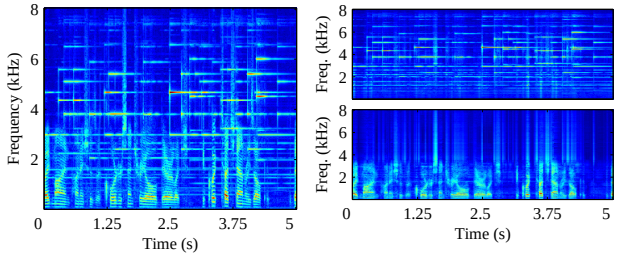

Fig. 2 Example of chime-speech separation: mixture (left), separated chime (top right) and separated speech (bottom right).

We tested three different settings: our algorithm by itself for 100 iterations; our algorithm after 500 iterations of the NMF algorithm; and finally our algorithm after 100 iterations of the original complex NMF. Each time, we computed the SNR of the separated sources. Our algorithm alone led to an improvement of +11.4 dB for the chime sounds and +9.71 dB for speech. After NMF converged and led to an improvement of +12.7 dB and +6.6 dB respectively, 50 iterations of our algorithm further improved the results to +13.1 dB and +9.8 dB. Finally, after complex NMF converged and led to improvements of +13.1 dB and +7 dB, 100 iterations of our algorithm led to +12 dB and +10 dB. The evolution of the various terms of (3) is shown in Fig. 1, and the final separation results in Fig. 2. The introduction of the consistency constraints after 100 iterations can be clearly seen. Altogether, we see that introducing the consistency constraints seemed to enable further improvements in the results obtained by NMF and complex NMF in terms of SNR on this task. We need to investigate their behavior more thoroughly in the future.

References

[1] J. Le Roux, Ph.D. dissertation, The University of Tokyo & Université Paris VI, Mar. 2009.

[2] H. Kameoka et al., *Proc. ICASSP*, Apr. 2009, pp. 3437–3440.

[3] D. W. Griffin and J. S. Lim, *IEEE Trans. ASSP*, vol. 32, no. 2, pp. 236–243, Apr. 1984.

[4] P. Smaragdis, *IEEE Trans. ASLP*, vol. 15, no. 1, pp. 1–12, Jan. 2007.